December 2017

End of Term 2016 Presidential Web Archive

Mark E. Phillips  
*University of North Texas, mark.phillips@unt.edu*

Kristy K. Phillips  
*University of North Texas, kristy.phillips@unt.edu*

Follow this and additional works at: [https://docs.lib.purdue.edu/atg](https://docs.lib.purdue.edu/atg)

Part of the Library and Information Science Commons

**Recommended Citation**  
DOI: [https://doi.org/10.7771/2380-176X.7874](https://doi.org/10.7771/2380-176X.7874)

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
of these titles in-house, including the foldouts, and any other large format titles. These scans will also be contributed to HathiTrust.

Another issue is public access to digitized content. As of this writing, HathiTrust continues to only allow the download of full-text items in the public domain, including U.S. government documents, as a member benefit. The public still has the ability to search across full-text documents, so finding that obscure quote from a hearing or report is possible. The public can view and download a document page by page, not the entire document as one file, which can be a major frustration if a user who is unaffiliated with any member institution needs a lengthy document. While this access model is a great improvement over no access at all, it is not what UC considers full access. The single-page download restriction prevents members of the public from fully engaging with any of the federal documents contained within the database, including the thousands UC and other libraries have and will continue to contribute. UC firmly believes that fully opening up federal documents to the public without restrictions aligns perfectly with HathiTrust’s mission to "contribute to...the common good by collaboratively collecting, organizing, preserving, communicating, and sharing the record of human knowledge." Being able to share these digitized documents openly and without restriction would also fulfill the UC Libraries’ mission to “provide[d] the broadest access to the world’s knowledge." UC is committed to working with and encouraging HathiTrust to remove the public download restrictions placed on federal documents, and we invite other HathiTrust members to do the same.

Working with HathiTrust has also been a great opportunity to brainstorm on various issues. UC and HathiTrust have been able to work through some of the issues both projects have encountered, such as reconciling various cataloging practices mentioned above. We have had preliminary discussions on resolving serials matching issues and identifying gaps in HathiTrust. A small example of gap filling: UC contributed several missing volumes of the Statistical Abstract of the United States, volumes that were non-destructively digitized so we may continue to retain the print for FedDocArc. UC and other digitization partners are also identifying publications we can target as priorities for digitization and inclusion in HathiTrust, such as titles from the FDLP's Essential Titles List.

In addition to HathiTrust, this undertaking has involved a number of players outside the University. UC signed a Shared Housing Agreement memorandum of understanding with the U.S. Government Publishing Office (GPO) in which provision of continued public access to the documents is explicitly spelled out. The California State Library, which oversees the FDLP in California, has been very supportive of the work we are doing to create a full collection of documents within the state. The State Library has allowed us some much-needed flexibility within the governing authority of the FDLP, so that we can work more efficiently to create the archive.

Next Steps

FedDocArc also requires UC to begin developing strategies to address several other issues. UC government documents librarians need to make some major decisions as FedDocArc moves forward. There are a number of questions to settle, such as which campuses will be responsible for contributing print publications to the archive and which campuses will contribute copies for digitization in the future. Will campuses split up the responsibility by agency, subject matter, or individual publications, or based on another option that has not yet been identified? What about CD-ROMs and other electronic materials, and born-digital content: how will these be included in FedDocArc? These are some examples of the questions remaining and the ongoing dialogue UC Libraries will need to continue among ourselves to resolve these issues.

The University of California Libraries are committed to completing the FedDocArc project and it has a great deal of support within the system. Having the collection digitized will open new avenues of discovery and research in scale and scope that had previously been unimaginable. FedDocArc has the potential to allow UC to open its collections to a large population outside the university, providing great public benefit, while at the same time retaining an archive of the print documents that will be preserved. FedDocArc is a project unlike anything the UC Libraries have done in the past, and we are looking forward to being able to share much of our collection of federal documents with the State of California, the nation, and the world via HathiTrust.

Endnotes


End of Term 2016 Presidential Web Archive

by Mark E. Phillips (Associate Dean for Digital Libraries, the University of North Texas) <Mark.Phillips@unt.edu>

and Kristy K. Phillips (University of North Texas) <kristy.phillips@unt.edu>

Introduction

During every Presidential election in the United States since 2008, a group of librarians, archivists, and technologists representing institutions across the nation can be found hard at work, preserving the federal web domain and documenting the changes that occur online during the transition.

Anecdotally, evidence exists that the data available on the federal web changes after each election cycle, either as a new president takes office, or when an incumbent president changes messages during their transition into a new term of office. Until 2004, nothing had been done to document this change. Originally, the National Archives and Records Administration (NARA) conducted the first large-scale capture of the federal web at the end of George W. Bush’s first term in office in 2004 (https://www.webharvest.gov/). This is noteworthy because, while institutions like the Library of Congress, the Government Publishing Office, and NARA itself have web archiving as part of their imperative, none of their mandates are so broad as to cover the capture and preservation of the entirety of the federal web. On April 15, 2008, NARA released the document “National Archives and Records Administration Web Harvesting Background Information,” which detailed the reasons why the organization decided not to continue this large-scale archival practice during the following election in 2008. As such, a group of interested organizations gathered together to continue the project.

The End of Term (EOT) projects began with the Internet Archive, the Library of Congress, the University of North Texas, the California Digital Library, and the U.S. Government Publishing Office working together to fill the void left by NARA and archive the entirety of the federal web during the transition period in the wake of the 2008 presidential election. Since that first capture, new partners have joined the team, including Harvard University in 2012, and George Washington University and Stanford University in 2016.
Every year, the process is updated and expanded. Every election brings its own challenges, but the unanticipated outcome of the presidential election of 2016 brought an especially eventful harvest, with people all over the country suddenly interested in what was captured during this particularly decisive transition. The EOT projects have several areas of organization, including seed collection, harvesting, and public outreach, that were affected by the changes brought by the most recent presidential election.

What to Harvest

The first step involved in a successful harvest is deciding what, exactly, needs to be captured. The End of Term project team has experimented with different ways of establishing the scope of the project each time it is completed, and several of them were used during the 2016 EOT project. Web harvesters require a set of starting URLs, or “seeds” that dictate where to begin the crawling process. To start, the harvester downloads the page designated by a seed URL, extracts all of the URLs on that page, then checks whether the extracted URLs have been crawled, and if they have not, it adds them to the list of URLs to crawl. This process is repeated until the list of new URLs has been exhausted, or until the crawler has been stopped by some other means. This can be done by the operator, or based on some threshold like total gigabytes downloaded, number of URLs in the crawl, or length of time crawling.

The federal web has a number of high-level websites that are entry points for users into the wide range of content that is available on the federal web. Sites like USA.gov provide an entry point in the format of a search and discovery portal. Unfortunately not all URLs in the federal web are identified in these systems, so the EOT project group first had to work to identify the overall scope of what content we would harvest. To identify the seed URLs that the EOT project would crawl, and therefore identify the scope of the crawling effort, the team used two primary methods of collecting seeds. These methods were bulk seed lists and URL nominations. These are both described in detail.

Bulk Seed Lists

It may be somewhat surprising, but there is not a definitive list of all of the domains and subdomains that are managed by the federal government. The U.S. General Services Administration (GSA) has created the U.S. Digital Registry which is an official list of a large number of these domains, but it is by no means exhaustive. Different groups within the government handle the registration of.gov and.mil domains, both of which are in the scope of the EOT project. Outside of the domain names, subdomains are often created and managed within the agency that created them, meaning they don’t make the standard lists of federal websites.

During the 2016 EOT project, the team used seven or eight different bulk seed lists, some from previous web crawls, and others collected from related projects. Once the lists were compiled, they were added to an instance of the URL Nomination Tool that the project team uses to manage them. Ultimately, a total of 43,674 seed URLs from ten different lists were added during the course of the project (http://digital2.library.unt.edu/nomination/eth2016_bulk).

URL Nominations

While domains and subdomains give broad targets for the EOT harvesters to crawl, there is important content that exists at all levels of an agency’s presence. This includes departmental, project, initiative, or committee home pages which often do not have their own subdomains. Of increasing importance are publications like PDFs, datasets, and other content-rich files which may not be discovered by the broader crawls that start out at higher levels of an agency’s website.

From the beginning, the team agreed it was important to allow people outside the interested organizations to submit government websites for themselves. This was the case again in 2016, and individuals were able to contribute to the project by submitting URLs to a new instance of the URL Nomination Tool for the websites they were interested in harvesting and preserving for the future. In addition to the URL, users were asked to include the branch of government, the specific government agency, and a title for their submission. The team received over 13,000 URLs nominated by 393 different nominators by the end of the 2016 project (http://digital2.library.unt.edu/nomination/eth2016/).

Social Media

During the prior harvest, the EOT project team realized that they were missing an important part of the government’s internet presence. Every day, many government agencies interact with and inform their constituents via social media sites like Facebook and Twitter. These interactions are also worth preserving as content of the federal web, and the team took steps to address that in 2012, and again in 2016. George Washington University was interested in using their locally-developed social media capture platform, Social Feed Manager, to accomplish the task, and they were responsible for collecting media from Twitter and Tumblr. The U.S. Digital Registry maintains an active list of the governmental social media accounts currently in use, and encourages agencies to register their accounts with these sites. This made data collection much easier. More than 9,000 social media accounts were targeted for collection during the 2016 EOT project.

continued on page 29
FTP Content

Many government agencies still use FTP (File Transfer Protocol) servers to disseminate reports, datasets, and other large collections of content. While the EOT project was originally only focused on HTTP-based content from the web, in 2016 the team expanded the project’s scope to include FTP content. The Internet Archive took responsibility for this portion of the project, and worked to capture all of the FTP content submitted during the nomination phase. This proved to be a difficult task, as the size and scope of the FTP content was much greater than expected. We found that there is a massive amount of content made available to the public via FTP servers from a wide assortment of federal agencies. The amount of content we captured from the FTP servers alone was larger than the entirety of the HTTP-based and social media content.

Harvesting the Content

The 2016 EOT project started in the middle of September, much as it has in prior years. Four separate institutions took responsibility for harvesting. The Internet Archive crawled the entirety of the bulk seed lists and the user-nominated content. The Library of Congress conducted crawls focused primarily on the legislative branch. The University of North Texas harvested the .mil domain, as well as the Department of Transportation and FEMA websites. George Washington University used its Social Feed Manager to harvest social media content.

The project team used the Open Source Heritrix Web Crawler for its harvesting activities, and saved all output as WARC (Web ARchival file format) files. The WARChive file format is an ISO (International Organization for Standardization) standard for storing content and HTTP transaction headers generated during the crawling process. Because all of the crawling partners used the same file format for storing archival web content, it was easy for us to share data between institutions.

Building a Collection of Publications

After looking through the URLs submitted via the URL Nomination Tool, the University of North Texas (UNT) decided that it would be a good idea to build a collection in the UNT Digital Library to house all of the PDF documents nominated directly. This highly-curated list of publications represents content that users were specifically interested in preserving, so UNT decided to offer item descriptions and easy access for these specific documents.

With this in mind, the project team at UNT created a collection called the End of Term Publications (https://digital.library.unt.edu/explore/collections/EOT/) and included over 1,900 PDF files in the collection. Volunteers created metadata for many of these items during the winter of 2016 and spring of 2017, which allowed UNT to make 60 percent of the documents with full descriptions available to the general public. Over 7,000 uses of the documents have been recorded to date. Many of these documents are focused on climate change and the environment, though parole forms and other documents from the Department of Justice and publications from the Department of Labor are also included in the collection.

Sharing the Harvested Content

In May of 2017, the project team began to compile all of the separately harvested data into a single location at the Internet Archive. In the past, the institutions involved in the project have used several technologies to transfer data, but for 2016 the team decided to go with something a little simpler, and shipped the data directly on large (8TB) hard drives. The data, stored in WARC files, included fixity hashes to verify file integrity. Altogether, the collecting partners gathered more than 200 TB of data. The Internet Archive loaded the aggregate collection of the 2016 EOT into an instance of the Wayback machine, and access records were added to the projects website (http://eotarchive.cdlib.org/).
Finally, the team suggested that the most helpful activity for volunteers was to nominate the URLs of the items that they believed most at risk via the URL Nomination Tool. This influx of nominations helped identify a wide range of content from websites to individual PDFs and datasets. It was a great help, and it allowed people to contribute in a way that they found meaningful. It also exposed a problem with the project: the team needed a better web presence to communicate with the public. Currently, the team has a Twitter account that was active during the project, but that is clearly not enough, as it is difficult to use as the only primary news and information outlet. In addition, the EOT project’s interface, which is hosted by the California Digital Library, wasn’t designed to have a section that listed new content, so updating the public via this resource simply wasn’t possible. Now, one of the major goals for the 2020 EOT project is to have a better news and information platform for communicating with the interested public.

**Conclusion**

The End of Term projects in 2008, 2012, and 2016 were volunteer efforts by a number of institutions across the U.S. The time, effort, and infrastructure are all donated by the participating organizations. The individuals from these institutions are the ones that moved the project forward and made it successful. The 2016 election cycle offered new challenges and opportunities in relation to project management, channeling user interest, fielding media requests, and gathering and sharing the harvested content. While there were challenges, they were insignificant in comparison to the overall benefit of the project, as well as the accomplishments of the project and its project team.

**References**

Data Refuge – https://www.datarefuge.org/
End of Term Archive Website – http://eotarchive.cdlib.org/
End of Term Publications Collection – https://digital.library.unt.edu/explore/collections/EOT/
URL Nomination Tool EOT 2016 Bulk – http://digital2.library.unt.edu/nomination/eth2016_bulk/